What data should I include in my POS tagging training set?

Zoey Liu[1] Masoud Jasbi[2] Christan Grant[1] Kenji Sagae[2] Emily Prud'hommeaux[3]

[1] University of Florida
[2] University of California, Davis
[3] Boston College

Motivation

Building training sets for understudied, endangered, and Indigenous languages faces **resource limitations** and **ethical constraints**.

Research Question: How can we construct effective POS tagging training sets when:

- Annotated data is scarce
- Manual annotation is expensive
- Data sharing may be ethically restricted

Why POS Tagging?

- Fundamental for language documentation
- Used in typological research, second language learning, pedagogical materials
- Data available across languages (Universal Dependencies)

Approach

We compared **three data-selection methods** across **60 languages** (112 treebanks, 12 families):

1. In-Context Learning (LLMs)

- **Model:** GPT-4.1-mini via API
- **Data:** 1,000 randomly sampled tokens as prompt examples
- **Cost:** ~$4 per language

2. Active Learning (AL)

- **Model:** Conditional Random Fields (CRF)
- **Strategy:** Uncertainty sampling - selects sentences with lowest confidence
- **Process:** Iteratively adds ~500 tokens per iteration
- **Initial set:** 1,000 tokens

3. Random Sampling (Baseline)

- Uniformly samples ~500 tokens per iteration
- Baseline for comparison with Active Learning

Data Source: Universal Dependencies v2.14

Results: In-Context Learning

Languages tested	60
F1 > 0.83	58 (97%)
F1 ≥ 0.90	~35 (58%)
Training tokens	1,000
Cost per language	$4

Performance Examples

- **French:** F1 = 0.97
- **English:** F1 = 0.93
- **Bulgarian:** F1 = 0.97
- **Hindi:** F1 = 0.90
- **Irish:** F1 = 0.90

Key Insight

For communities where **data sharing via API is ethical and acceptable**, LLMs provide excellent first-pass performance with **minimal annotation cost**.

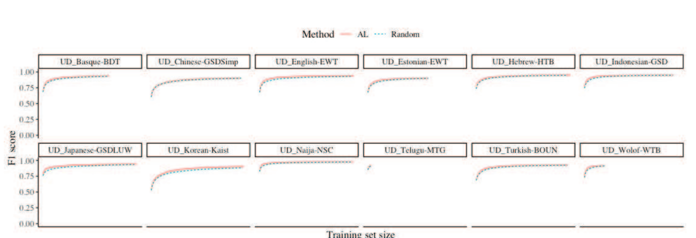

Figure 1: F1 scores across diverse language families

Results: Active Learning

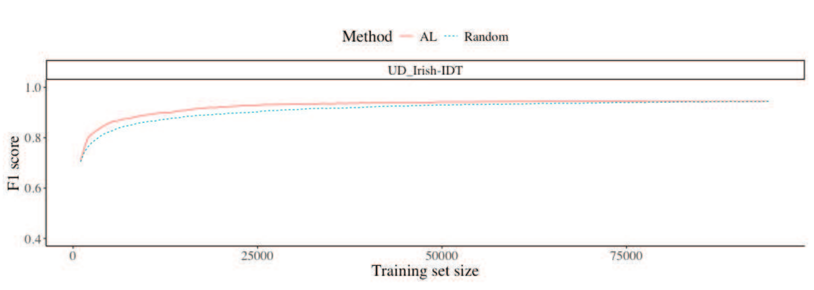

Figure 2: Irish (endangered, EGIDS 6b): AL vs Random Sampling

Key Observations

- **Rapid growth:** F1 increases quickly until 4,500-5,500 tokens
- **Irish example:**
  - 1,000 tokens → F1 = 0.71
  - 4,500 tokens → F1 = 0.85
  - 12,000 tokens → F1 = 0.90
- **Plateau:** Performance stabilizes after ~20,000 tokens

Statistical Validation

Use LLMs to bootstrap low-resource language documentaton.

GPT-4.1-mini with 1,000 tokens achieves F1 > 0.83 in 58/60 languages

Active Learning learns 2x faster than random sampling

Sweet spot: 4,500-5,500 tokens for reasonable F1 with Active Learning

We used **Bayesian growth curve modeling** to quantify learning speed:

$$F1 = \alpha - (\alpha - \beta)e^{-(\gamma t)^{\delta}}$$

Parameters:

- α = upper asymptote (max F1)
- β = lower asymptote (starting F1)
- γ = **growth rate** (learning speed)
- δ = shape parameter

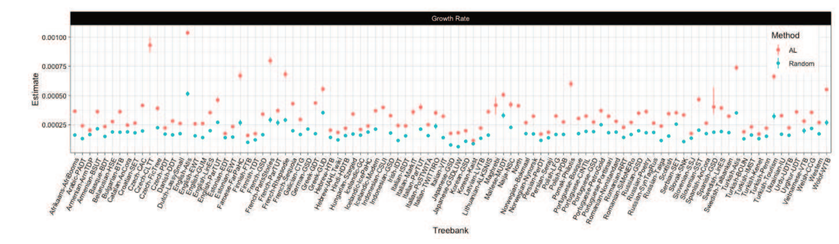

Figure 3: Growth rates: AL consistently faster than random sampling

Finding

Active Learning reaches target F1 scores approximately 2x faster than random sampling across all 112 treebanks.

What Affects Tag Performance?

Mixed-effects regression on individual POS tag F1 scores:

Factor	Effect	Interpretation
Word Entropy	β = +0.036**	More diverse vocabulary → Better
Syntax Entropy	β = −0.05**	More diverse syntax → Worse
Tag Probability	β = −0.17 (n.s.)	No significant effect

KL-Divergence Analysis

- Training sets **more similar to test distribution** → higher F1
- As training size ↑, KL-divergence ↓
- **Negative correlation:** KL-divergence vs F1 (β = -5.35, $p < 0.001$)

Recommendations

Scenario	Strategy	Training Size	Effort
Data sharable via API	GPT-4.1-mini	1,000 tokens	Low cost (~$4)
Data must remain local	Active Learning	4,500-5,500 tokens	Moderate (2x faster)
Highly restricted / Maximum accuracy	Random Sampling	20,000+ tokens	High (slower convergence)

Data Sovereignty Matters

Many Indigenous communities require **data to remain within community control**, making Active Learning the **ethical and effective choice**.

Conclusions

For Ethical Data Sharing

LLMs can bootstrap annotation with minimal data:

- Only 1,000 tokens needed
- F1 > 0.83 in 97% of tested languages
- Cost-effective (~$4 per language)

For Data Sovereignty

Active Learning maximizes efficiency while respecting community values:

- Reaches F1 > 0.85 with 4,500-5,500 tokens
- 2x faster learning than random sampling
- Data remains under community control

Methodological Contribution

- **First large-scale cross-linguistic AL study** with statistical validation
- **Growth curve modeling** provides rigorous quantification
- Framework applicable to other low-resource NLP tasks

Contact: christan@ufl.edu | liu.ying@ufl.edu

Code: github.com/ufcompling/unlabeled_pos

Paper: ACL Anthology 2025.findings-emnlp.448

QR Code: Scan for full project details at ufdatastudio.com

DATA STUDIO

UF UNIVERSITY of FLORIDA